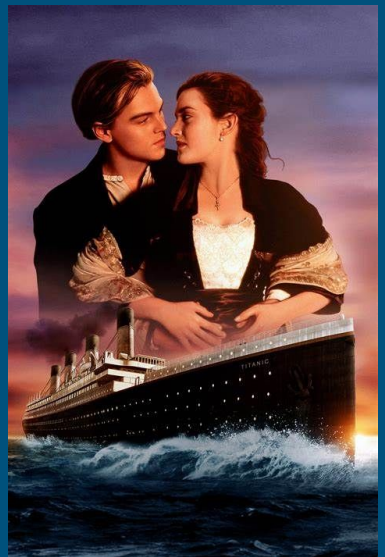
An Analysis of the Titanic Movie



Wenhe Chen, Angel Wang, Miao Zhang, Yihan Zhang

Research Objectives

- Film review emotion analysis
- Film review trends forecast
- Lexical analysis of film review

Preparation of data set

Data collection: the data is collected using a Selenium web driver to automatically navigate and scrape IMDb movie reviews.

Data cleaning: data quality assessment, data integration, data conversion, and etc

Data variables: Title, Author, Date, Review

Algorithm Selection

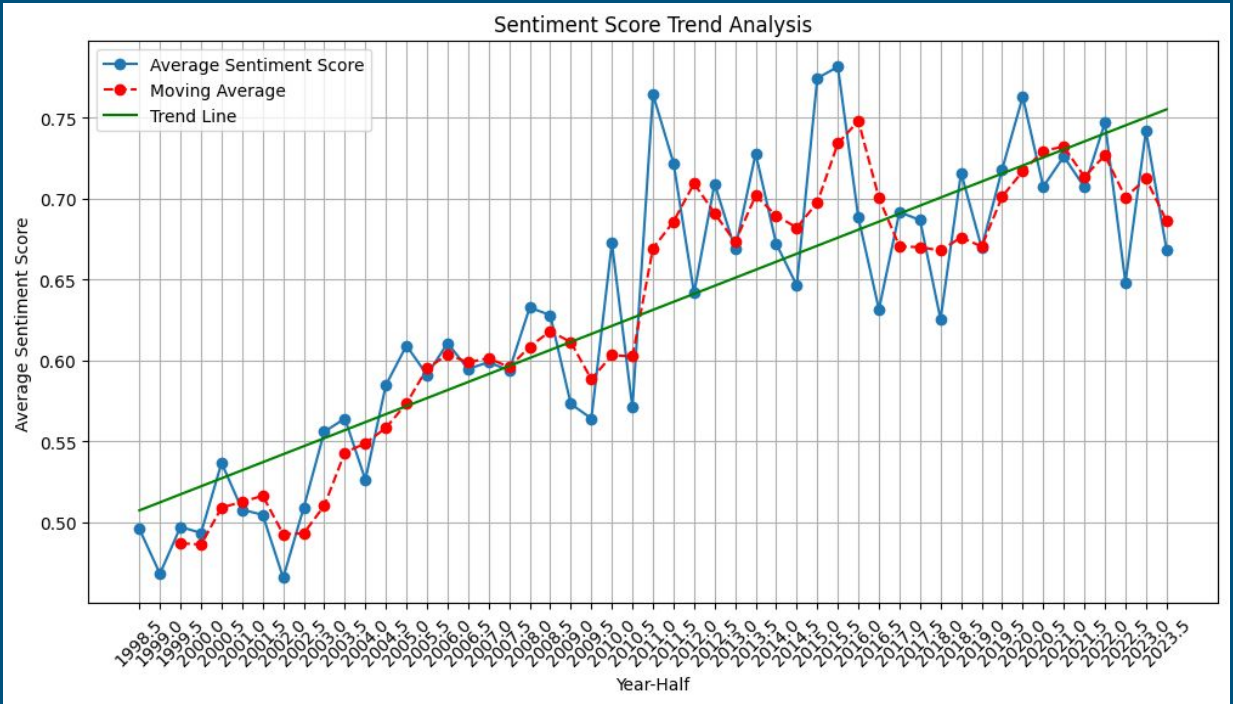
Transformer VS RNN, LSTM, Naive Bayes, SVM, etc:

less efficient when dealing with long-distance dependencies and take longer to train / poor at processing complex textual data and capturing subtle emotional differences

Hugging Face Transformers

- Pre-training: Models in the library are first pre-trained on a large corpus of texts. This stage aims for the model to learn the semantic representation of text, capturing the relationships between different words through self-attention mechanisms.
- Fine-tuning: After pre-training, models can be fine-tuned to adapt to specific NLP tasks, such as text classification or sentiment analysis. The fine-tuning phase involves retraining the model with task-specific data sets to adapt it to a particular job.
- Feature extraction: Users can use the library's API to extract text features from pre-trained models that can be used for various NLP tasks.
- Applications: Users can apply fine-tuned models to various text-related tasks, such as question answering, text generation, machine translation, and more.
- Hugging Face's Transformers library not only implements basic Transformer models, but also includes a variety of pre-trained models such as BERT, GPT, T5, etc., which demonstrate excellent performance on specific NLP tasks.

Trends since its release on IMDB



Half a year is an interval.

Take the average of six months of sentiment analysis scores.

.0 means first half of the year

.5 means the second half of the year

Predict future trends

	Year-Half_Numeric	Predicted_Sentiment_Score	Lower_Bound	Upper_Bound
0	2024.0	0.713079	0.614529	0.811629
1	2024.5	0.740517	0.638499	0.842535
2	2025.0	0.721350	0.615977	0.826722
3	2025.5	0.722681	0.614068	0.831295
4	2026.0	0.755096	0.643359	0.866833
5	2026.5	0.742622	0.627846	0.857398
6	2027.0	0.744595	0.626707	0.862483
7	2027.5	0.766159	0.645348	0.886970
8	2028.0	0.754468	0.630803	0.878133
9	2028.5	0.743119	0.616508	0.869730

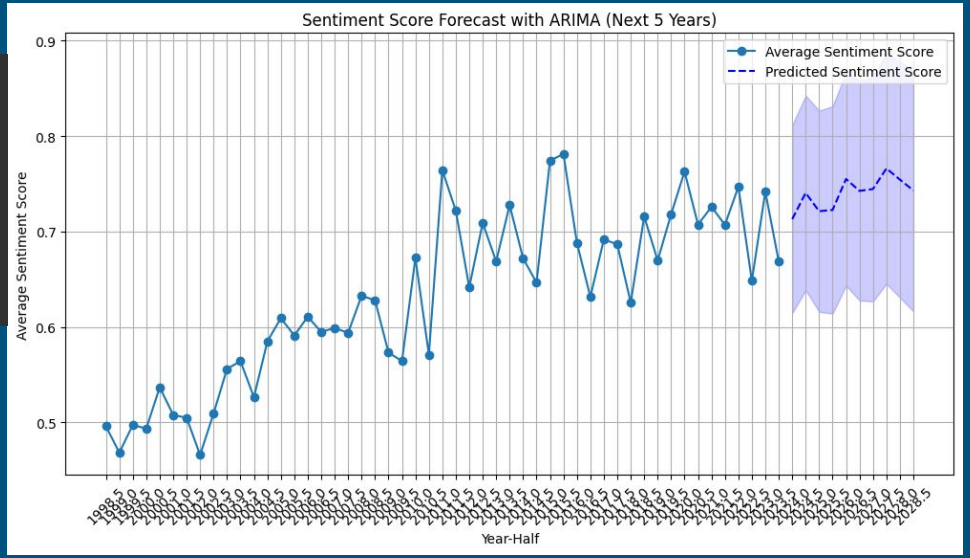
```python
import itertools
from statsmodels.tsa.stattools import arma_order_select_ic

p = d = q = range(0, 3)
pdq = list(itertools.product(p, d, q))

# Assuming 'Sentiment_Score' is your target variable
results_aic = []
for order in pdq:
    try:
        model = SARIMAX(score_by_time['Sentiment_Score'], order=order)
        results = model.fit(disp=False)
        results_aic.append((order, results.aic))
    except:
        continue

# Find the order with the lowest AIC
best_order = min(results_aic, key=lambda x: x[1])[0]
print(f"Best ARIMA Order: {best_order}")
```

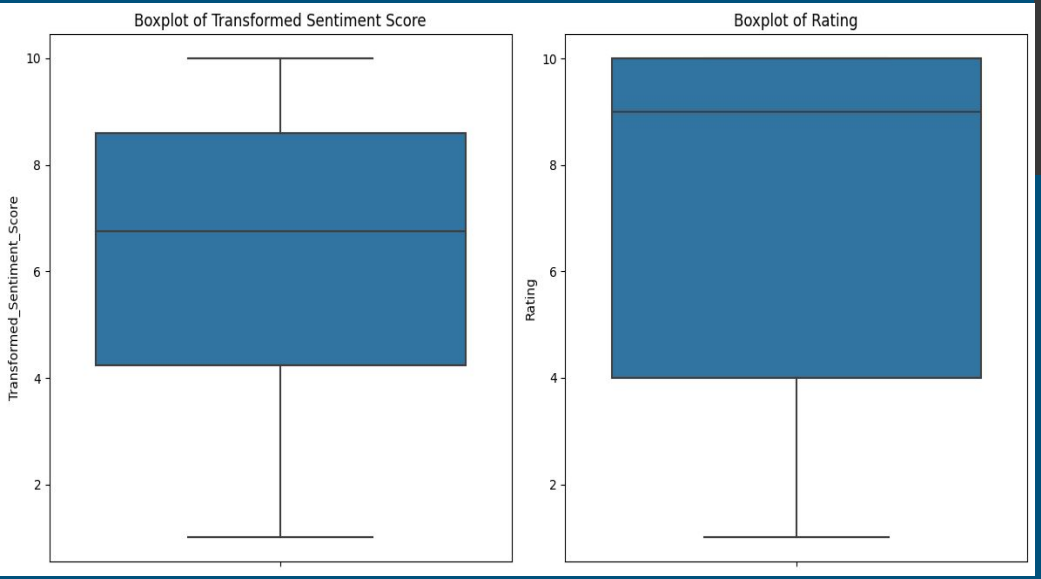
```
Best ARIMA Order: (0, 1, 1)
```



Predicting by ARIMA (0,1,1) (95% confidence interval)

Conclusion of discovery 1

Transformed Sentiment Score VS Rating:




```
Correlation Matrix:
                            Transformed_Sentiment_Score    Rating
Transformed_Sentiment_Score                     1.000000  0.757396
Rating                                          0.757396  1.000000

Spearman Correlation Coefficient:  0.7055647694706387
Spearman P-value:  2.1264551409578257e-181

Descriptive Statistics:
       Transformed_Sentiment_Score       Rating
count                  1200.000000  1200.000000
mean                      6.383814     7.029167
std                       2.624475     3.484983
min                       1.002166     1.000000
25%                       4.233460     4.000000
50%                       6.760628     9.000000
75%                       8.589163    10.000000
max                       9.998989    10.000000
```

Transformed Sentiment scores and ratings are significant positively correlated, but not always consistent.

The side highlights the limitations of sentiment analysis, the subjectivity of user ratings, cultural differences, and etc

Word Frequency Analysis

negative

```
[('cameron', 144),
 ('effect', 133),
 ('character', 115),
 ('sink', 97),
 ('hour', 94),
 ('special', 93),
 ('jack', 87),
 ('kate', 73),
 ('scene', 71),
 ('winslet', 69),
 ('dicaprio', 69),
 ('class', 65),
 ('rise', 64),
 ('real', 63),
 ('oscar', 62),
 ('acting', 61),
 ('plot', 60),
 ('life', 57),
 ('hate', 55),
 ('girl', 55)]
```

positive

```
[('cameron', 596),
 ('jack', 566),
 ('rise', 467),
 ('kate', 455),
 ('winslet', 435),
 ('dicaprio', 367),
 ('life', 336),
 ('character', 327),
 ('leonardo', 326),
 ('feel', 309),
 ('scene', 297),
 ('heart', 281),
 ('romance', 253),
 ('effect', 249),
 ('amazing', 240),
 ('beautiful', 236),
 ('actor', 228),
 ('real', 204),
 ('end', 195),
 ('performance', 187)]
```

We choose sentimental scores >0.8 (>8/10) for positive review analysis, and sentimental scores<0.3 (<3/10) for negative review analysis.

Trends in Reviews analysis of Top Frequency Words

Negative

```
Top words for 1998-2002: [('effect', 92), ('character', 83), ('cameron', 82), ('special', 68), ('hour', 50)]
Top words for 2003-2007: [('cameron', 33), ('jack', 30), ('rise', 29), ('sink', 29), ('kate', 27)]
Top words for 2008-2012: [('action', 11), ('effect', 10), ('real', 10), ('awful', 10), ('sink', 9)]
Top words for 2013-2017: [('cameron', 11), ('scene', 7), ('awful', 4), ('kate', 4), ('long', 4)]
Top words for 2018-2022: [('jack', 20), ('life', 14), ('star', 14), ('sink', 12), ('rise', 11)]
```

In the early days of a film, audiences paid more attention to novel special effects and character development, which were the main criteria for evaluating the film. Over time, however, audiences began to evaluate the film more deeply and pay more attention to the style of the director, the depth of the story, and the complexity of the characters.

Positive words Frequency

```
Top words for 1998-2002: [('cameron', 134), ('winslet', 87), ('kate', 85), ('dicaprio', 72), ('jack', 65)]
Top words for 2003-2007: [('jack', 148), ('kate', 118), ('winslet', 115), ('cameron', 115), ('rise', 108)]
Top words for 2008-2012: [('jack', 114), ('cameron', 107), ('rise', 96), ('winslet', 71), ('kate', 65)]
Top words for 2013-2017: [('jack', 101), ('rise', 98), ('cameron', 59), ('kate', 58), ('character', 52)]
Top words for 2018-2022: [('cameron', 115), ('kate', 101), ('jack', 86), ('winslet', 86), ('leonardo', 79)]
```

Over time, character nouns and actors' names appear more frequently in positive words, which may reflect the audience's emotional connection to the characters and continued recognition of the actors' performances.

Potential Business Value

- The reproduction of classic event type films
- Improving products or services: Consumer preference insights, segmentation strategies based on group subjectivity
- Risk management: Timely detection of negative sentiment and low ratings in comments can serve as a crisis warning. Companies can manage their brand reputation more effectively by analyzing public sentiment and evaluation
- Analyzing trends in sentiment scores and ratings can help predict market trends and consumer behavior, providing a basis for strategic decisions
- By comparing the emotional scores and ratings of different movies or products, you can better understand the strengths and weaknesses of your competitors

Reference

"Titanic Reviews & Ratings." IMDb, IMDb.com, 6 Dec. 2023, https://www.imdb.com/title/tt0120338/reviews.

Hugging Face. "Transformers." Hugging Face, 2023, https://huggingface.co/transformers.